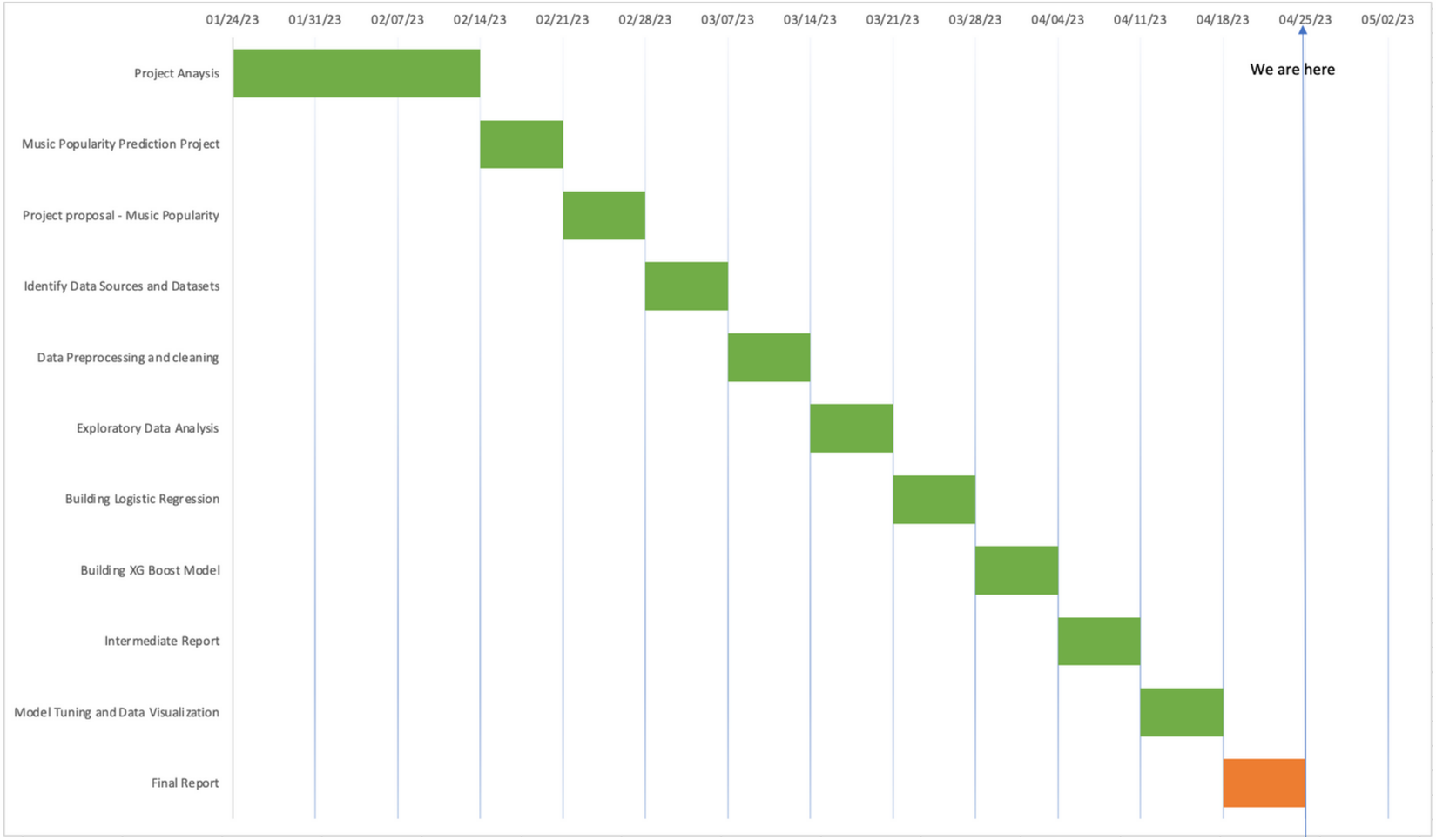
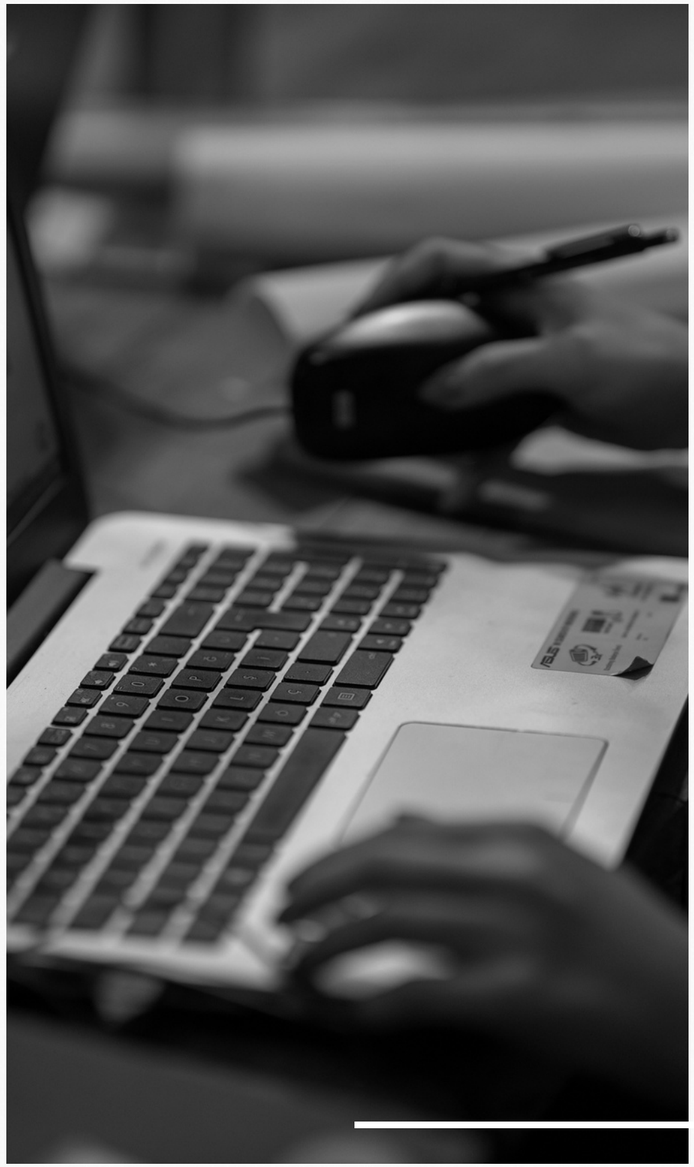
MUSIC POPULARITY PREDICTION

Group 5:
Anusha Muniraju
Chieh-Hsin Wu
Huyen Nguyen khanh tran
Shiva Prasad Reddy
Sri Harsha Somayajula
Surabhi Suresh

Execution Plan



BUSINESS PROBLEM



One of the most critical challenges for music investors is accurately valuing new artists and effectively managing established singers who have sold their song catalogs for significant sums of money in recent years. However, the music industry is highly unpredictable, and there is no guarantee that every song from an established singer will be a hit. Therefore, predicting a song's popularity is crucial for decision-makers in the industry. By leveraging data analysis and market research, music investors can make informed decisions about which artists and songs to invest in, ultimately maximizing their return on investment.

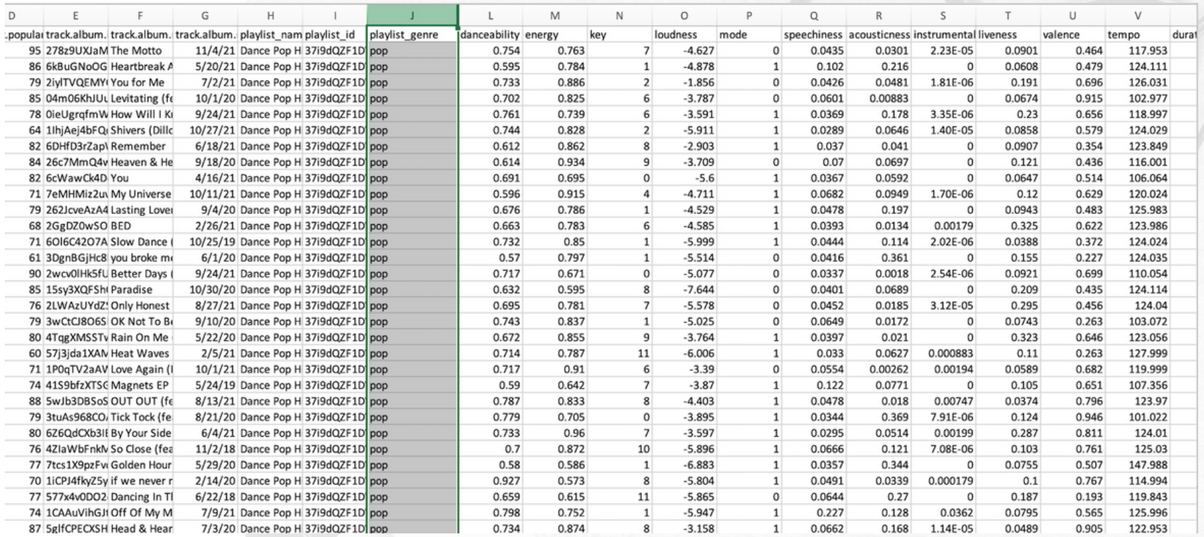
We will answer these questions:

- What specific features or qualities do successful songs tend to possess?
- What are the primary factors that drive a song's success, and which ones have the greatest impact?
- Can the popularity of past songs be used to predict the success of new ones?

OUR GOAL

WE AIM TO INTRODUCE A MODEL THAT CAN FORECAST THE LIKELIHOOD OF A SONG'S SUCCESS, AS DEFINED BY ITS INCLUSION IN THE TOP 40% OF BILLBOARD'S CHART.

DATA DESCRIPTION



D	E	F	G	H	I	J	L	M	N	O	P	Q	R	S	T	U	V	
.popula	track.album.	track.album.	track.album.	playlist_nam	playlist_id	playlist_genre	danceability	energy	key	loudness	mode	speechiness	acousticness	instrumental	liveness	valence	tempo	durat
95	278z9UXJaM	The Motto	11/4/21	Dance Pop H	37i9dQZF1D	pop	0.754	0.763	7	-4.627	0	0.0435	0.0301	2.23E-05	0.0901	0.464	117.953	
86	6kBuGNoOG	Heartbreak A	5/20/21	Dance Pop H	37i9dQZF1D	pop	0.595	0.784	1	-4.878	1	0.102	0.216	0	0.0608	0.479	124.111	
79	2iyITVQEMY	You for Me	7/2/21	Dance Pop H	37i9dQZF1D	pop	0.733	0.886	2	-1.856	0	0.0426	0.0481	1.81E-06	0.191	0.696	126.031	
85	04m06KhJUu	Levitating (fe	10/1/20	Dance Pop H	37i9dQZF1D	pop	0.702	0.825	6	-3.787	0	0.0601	0.00883	0	0.0674	0.915	102.977	
78	0ieUgrqfmW	How Will I Kı	9/24/21	Dance Pop H	37i9dQZF1D	pop	0.761	0.739	6	-3.591	1	0.0369	0.178	3.35E-06	0.23	0.656	118.997	
64	1lhjAej4bFQ	Shivers (Dillc	10/27/21	Dance Pop H	37i9dQZF1D	pop	0.744	0.828	2	-5.911	1	0.0289	0.0646	1.40E-05	0.0858	0.579	124.029	
82	6DHfD3rZap	Remember	6/18/21	Dance Pop H	37i9dQZF1D	pop	0.612	0.862	8	-2.903	1	0.037	0.041	0	0.0907	0.354	123.849	
84	26c7MmQ4v	Heaven & He	9/18/20	Dance Pop H	37i9dQZF1D	pop	0.614	0.934	9	-3.709	0	0.07	0.0697	0	0.121	0.436	116.001	
82	6cWawCk4D	You	4/16/21	Dance Pop H	37i9dQZF1D	pop	0.691	0.695	0	-5.6	1	0.0367	0.0592	0	0.0647	0.514	106.064	
71	7eMHMiz2uı	My Universe	10/11/21	Dance Pop H	37i9dQZF1D	pop	0.596	0.915	4	-4.711	1	0.0682	0.0949	1.70E-06	0.12	0.629	120.024	
79	262JcveAzA4	Lasting Lover	9/4/20	Dance Pop H	37i9dQZF1D	pop	0.676	0.786	1	-4.529	1	0.0478	0.197	0	0.0943	0.483	125.983	
68	2GgDZ0wSO	BED	2/26/21	Dance Pop H	37i9dQZF1D	pop	0.663	0.783	6	-4.585	1	0.0393	0.0134	0.00179	0.325	0.622	123.986	
71	6OI6C42O7A	Slow Dance (	10/25/19	Dance Pop H	37i9dQZF1D	pop	0.732	0.85	1	-5.999	1	0.0444	0.114	2.02E-06	0.0388	0.372	124.024	
61	3DgnBGjHc8	you broke me	6/1/20	Dance Pop H	37i9dQZF1D	pop	0.57	0.797	1	-5.514	0	0.0416	0.361	0	0.155	0.227	124.035	
90	2wcv0lHk5fU	Better Days (	9/24/21	Dance Pop H	37i9dQZF1D	pop	0.717	0.671	0	-5.077	0	0.0337	0.0018	2.54E-06	0.0921	0.699	110.054	
85	15sy3XQFSh	Paradise	10/30/20	Dance Pop H	37i9dQZF1D	pop	0.632	0.595	8	-7.644	0	0.0401	0.0689	0	0.209	0.435	124.114	
76	2LWAzUYdZ	Only Honest	8/27/21	Dance Pop H	37i9dQZF1D	pop	0.695	0.781	7	-5.578	0	0.0452	0.0185	3.12E-05	0.295	0.456	124.04	
79	3wCtCJ8O6S	OK Not To Be	9/10/20	Dance Pop H	37i9dQZF1D	pop	0.743	0.837	1	-5.025	0	0.0649	0.0172	0	0.0743	0.263	103.072	
80	4TqgXMSSTv	Rain On Me	5/22/20	Dance Pop H	37i9dQZF1D	pop	0.672	0.855	9	-3.764	1	0.0397	0.021	0	0.323	0.646	123.056	
60	57j3jda1XAN	Heat Waves	2/5/21	Dance Pop H	37i9dQZF1D	pop	0.714	0.787	11	-6.006	1	0.033	0.0627	0.000883	0.11	0.263	127.999	
71	1P0qTV2aAV	Love Again (I	10/1/21	Dance Pop H	37i9dQZF1D	pop	0.717	0.91	6	-3.39	0	0.0554	0.00262	0.00194	0.0589	0.682	119.999	
74	41S9bfzXTSC	Magnets EP	5/24/19	Dance Pop H	37i9dQZF1D	pop	0.59	0.642	7	-3.87	1	0.122	0.0771	0	0.105	0.651	107.356	
88	5wJb3DBSoS	OUT OUT (fe	8/13/21	Dance Pop H	37i9dQZF1D	pop	0.787	0.833	8	-4.403	1	0.0478	0.018	0.00747	0.0374	0.796	123.97	
79	3tuAs968CO	Tick Tock (fe	8/21/20	Dance Pop H	37i9dQZF1D	pop	0.779	0.705	0	-3.895	1	0.0344	0.369	7.91E-06	0.124	0.946	101.022	
80	6Z6QdCXb3I	By Your Side	6/4/21	Dance Pop H	37i9dQZF1D	pop	0.733	0.96	7	-3.597	1	0.0295	0.0514	0.00199	0.287	0.811	124.01	
76	4ZIaWbFnkN	So Close (fea	11/2/18	Dance Pop H	37i9dQZF1D	pop	0.7	0.872	10	-5.896	1	0.0666	0.121	7.08E-06	0.103	0.761	125.03	
77	7tcs1X9pzFw	Golden Hour	5/29/20	Dance Pop H	37i9dQZF1D	pop	0.58	0.586	1	-6.883	1	0.0357	0.344	0	0.0755	0.507	147.988	
70	1iCPJ4fkyZ5y	if we never r	2/14/20	Dance Pop H	37i9dQZF1D	pop	0.927	0.573	8	-5.804	1	0.0491	0.0339	0.000179	0.1	0.767	114.994	
77	577x4v0DO2	Dancing In Tl	6/22/18	Dance Pop H	37i9dQZF1D	pop	0.659	0.615	11	-5.865	0	0.0644	0.27	0	0.187	0.193	119.843	
74	1CAAuVihGJ	Off Of My M	7/9/21	Dance Pop H	37i9dQZF1D	pop	0.798	0.752	1	-5.947	1	0.227	0.128	0.0362	0.0795	0.565	125.996	
87	5glfCPECXSH	Head & Hear	7/3/20	Dance Pop H	37i9dQZF1D	pop	0.734	0.874	8	-3.158	1	0.0662	0.168	1.14E-05	0.0489	0.905	122.953	

With the Spotify API, we have access to a vast amount of music data from Spotify, making it an excellent tool for building various systems.

We're utilizing the Spotify API to extract song features that IS integrated into our project.

1 Number of Observations - 34,080

2 6 genres - rock, rap, r&b, pop, latin, edm

Each genre has about 5500 observations

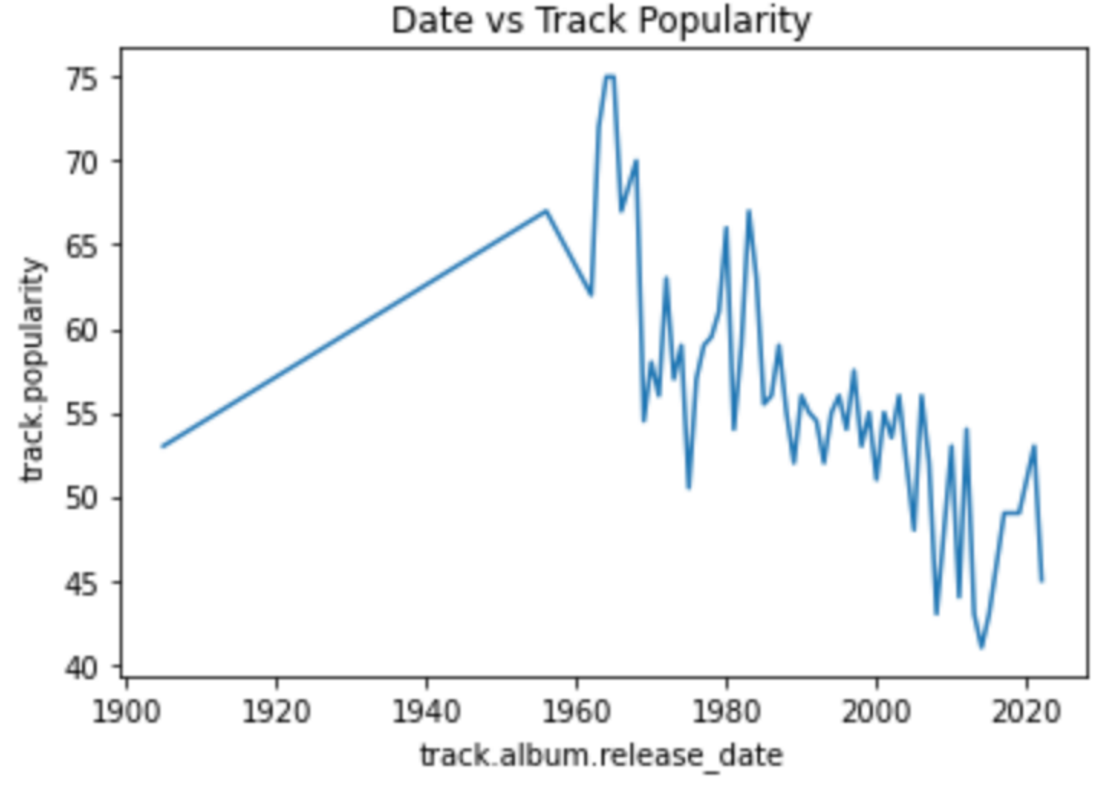
3

Variables	23
Numerical	14
Categorical	9

4 Target variable - Popularity

Type: quantitative - discrete

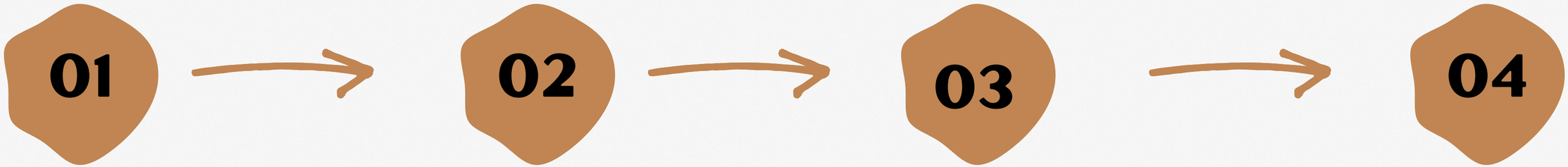
5 Time period: 1956-2022



- THE NUMBER OF SONGS BEING RELEASED INCREASES, MAKING IT HARDER FOR ANY INDIVIDUAL SONG TO GAIN AS MUCH POPULARITY AS SONGS IN THE PAST.

- THERE MAY BE CHANGES IN MUSICAL TASTES AND PREFERENCES OVER TIME THAT ALSO AFFECT THE POPULARITY OF SONGS.

Exploratory Data Analysis



CHECK FOR N/A AND 0 VALUE

CHECK FOR INCONSISTENCY IN TRACK.ALBUM_RELEASE_DATE COLUMN

CHECK POPULARITY'S DISTRIBUTION

CHECK MULTICOLLINEARITY BETWEEN INDEPENDENT VARIABLES

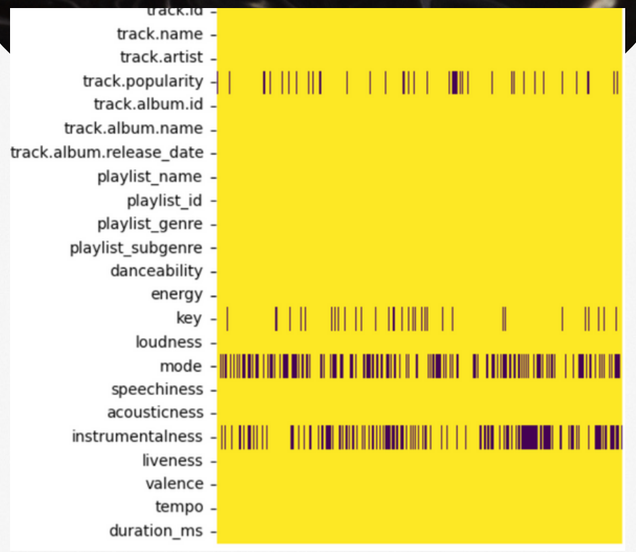
DATA CLEANING

1. Checking 0 value:

   - Key: Acceptable because 0 means C
   - Mode: Acceptable because of its domain (0,1)
   - Instrumentalness: Acceptable because the closer the instrumentalness value is to 0, the greater likelihood the track contains vocal content.

2. Checking for NA values

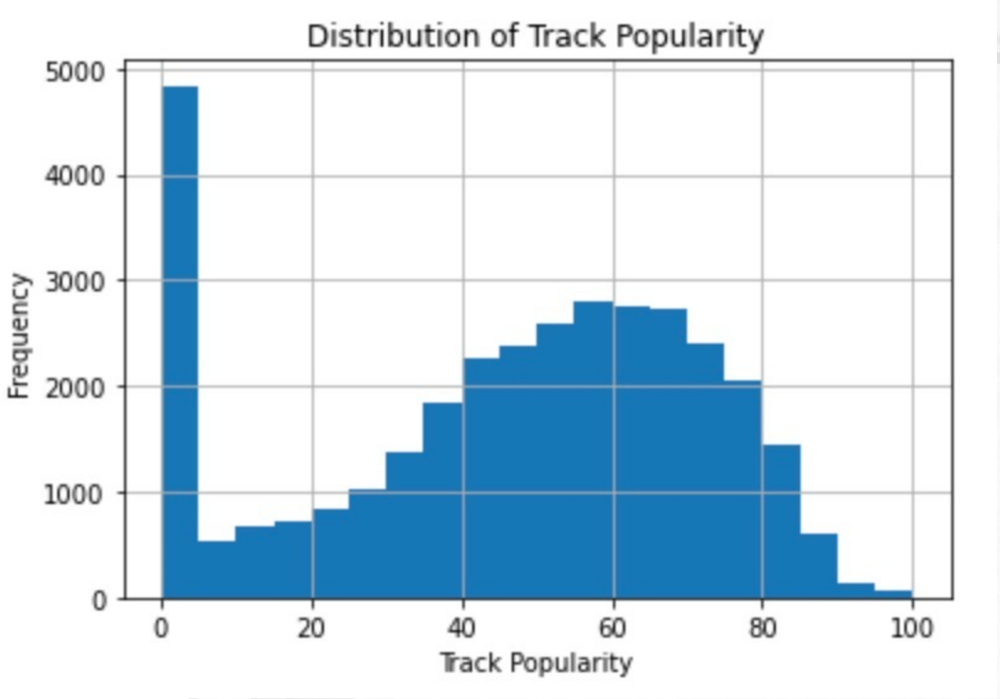
3. Inconsistent Date formats



1/1/81	1981
1981	1905
6/1/78	1978
1981	1905
1/1/75	1975
5/14/10	2010
10/15/86	1986
1975	
1976	

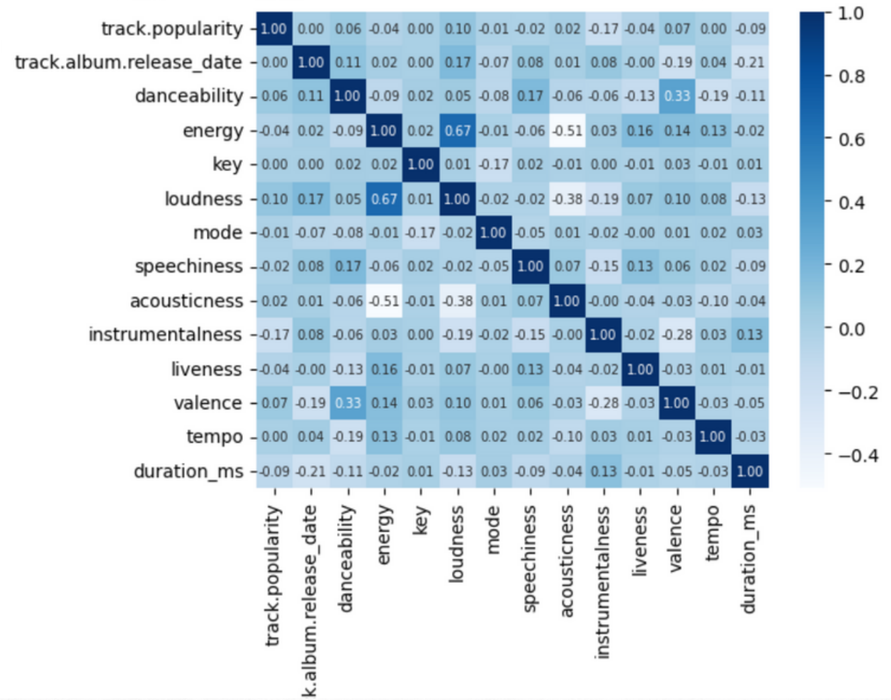
DATA ANALYSIS

Distribution of TARGET variable

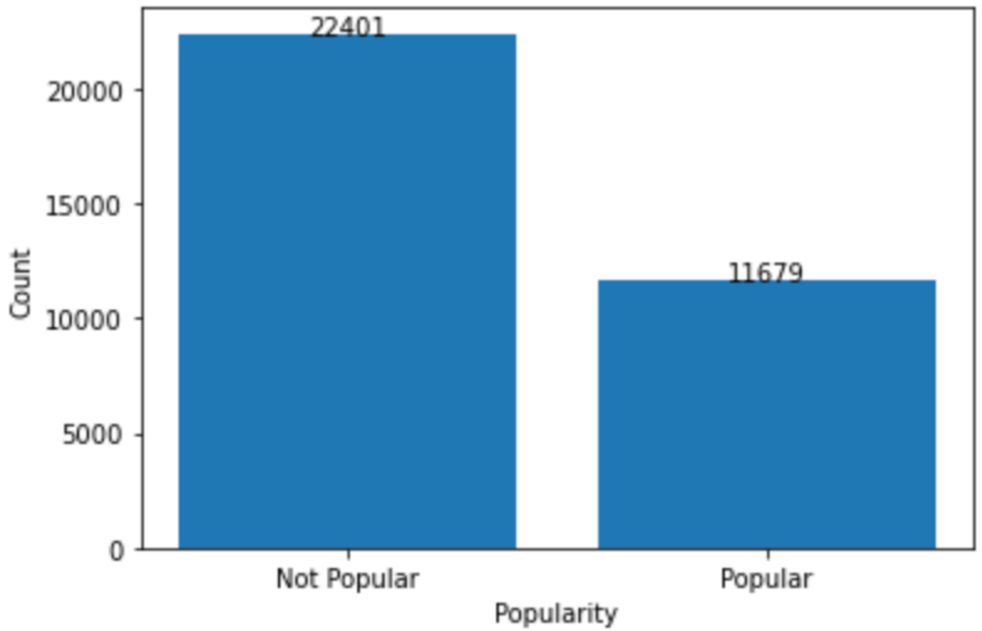


- Popularity scores are skewed
- Most of the songs have popularity score <60
- The dataset is quite unbalanced

Correlation Matrix



- High correlation between loudness and energy
- Not much correlation b/w independent variables and popularity.



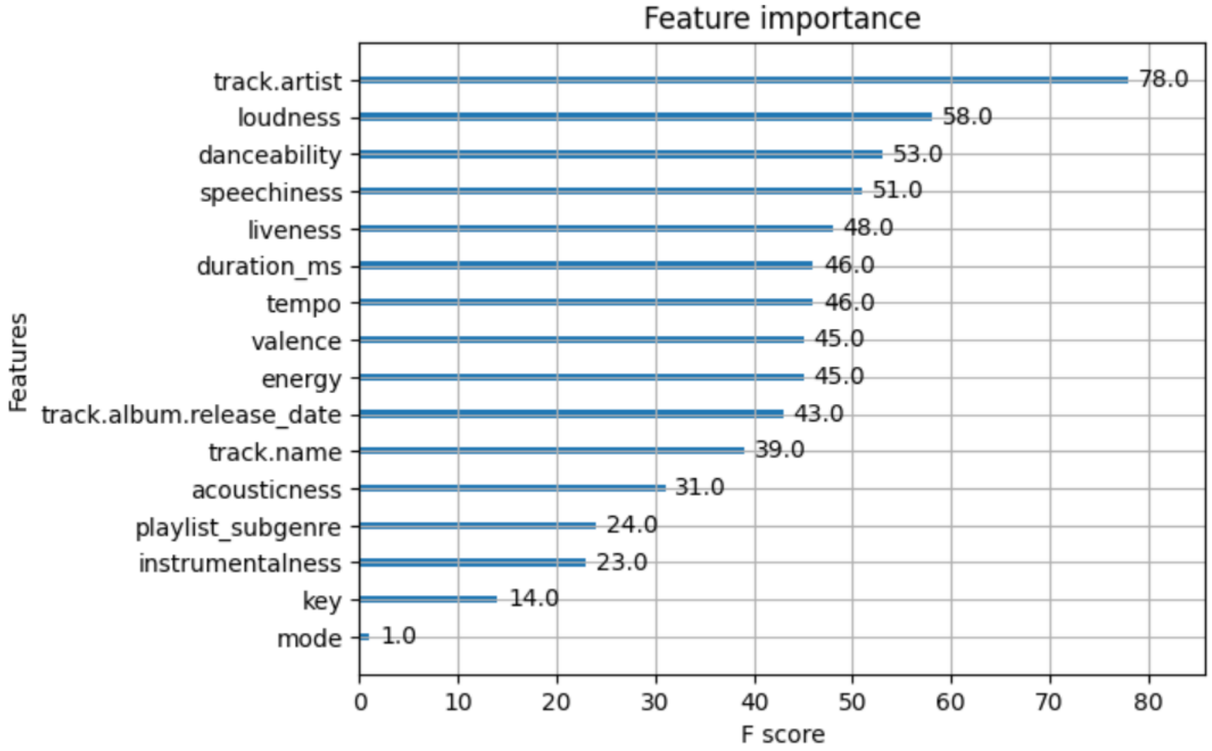
- For the purpose of our analysis we encode the popularity column which originally contains values from 0-100
- Categorize popularity to 0 and 1

  Popularity >= 60 ----> " 1"

  Popularity < 60 ----> " 0"
- The resulting dataset has a column popularity_encoded with 22401 rows having value 0 and 11679 rows having value 1

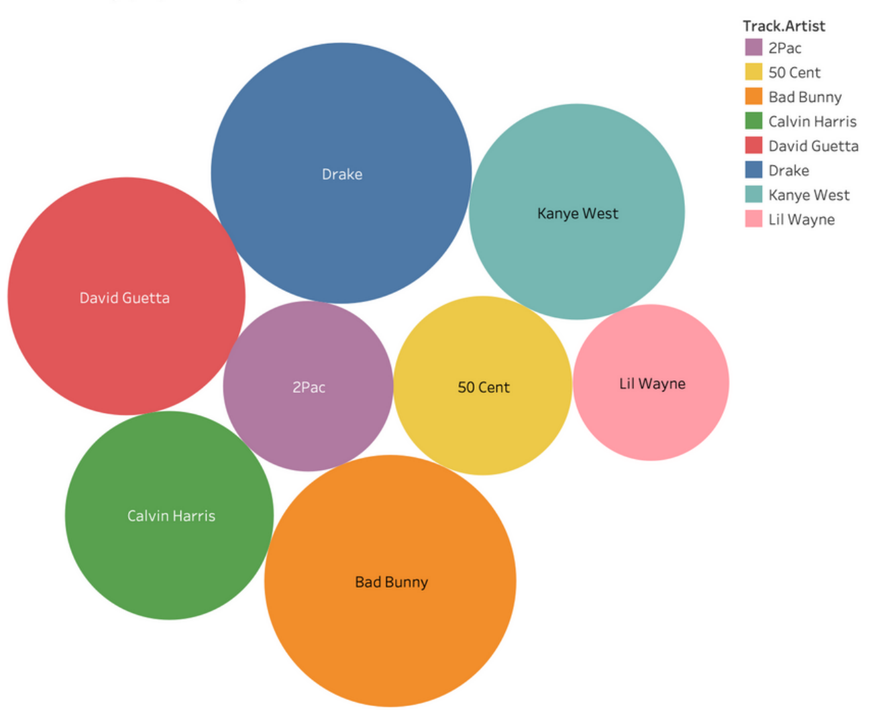
Feature Analysis

- The F-score is calculated as the number of times a feature is split on across all trees in the model, weighted by the number of training samples that go through those splits.
- The higher the F-score of a feature, the more important it is in determining the target variable.
- Therefore, the features with higher F-scores are considered to be more important for the model's prediction.



Popular Artists

Feature analysis shows that ARTIST, SPEECHINESS, DURATION, DANCEABILITY, LOUDNESS and LIVENESS play a very important role in determining the success of a song



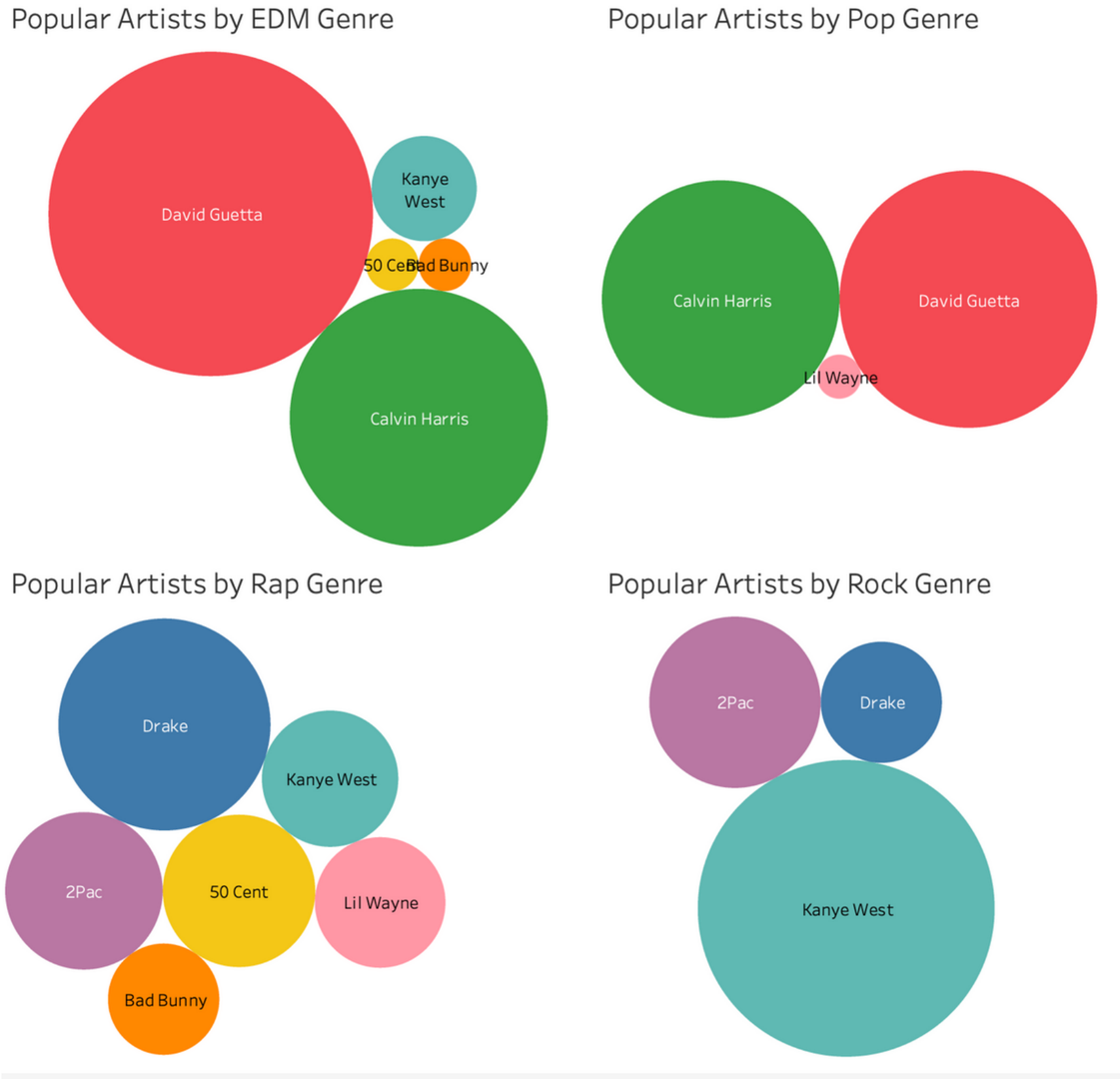
- Drake is by far the most popular artist.

TOP5 POPULAR ARTISTS PERCENTAGE

```
R&B   - 14.38%
ROCK  - 14.08%
POP   - 13.97%
RAP   - 11.77%
EDM   - 10.91%
LATIN - 8.76%
```

Famous Artists in each GENRE

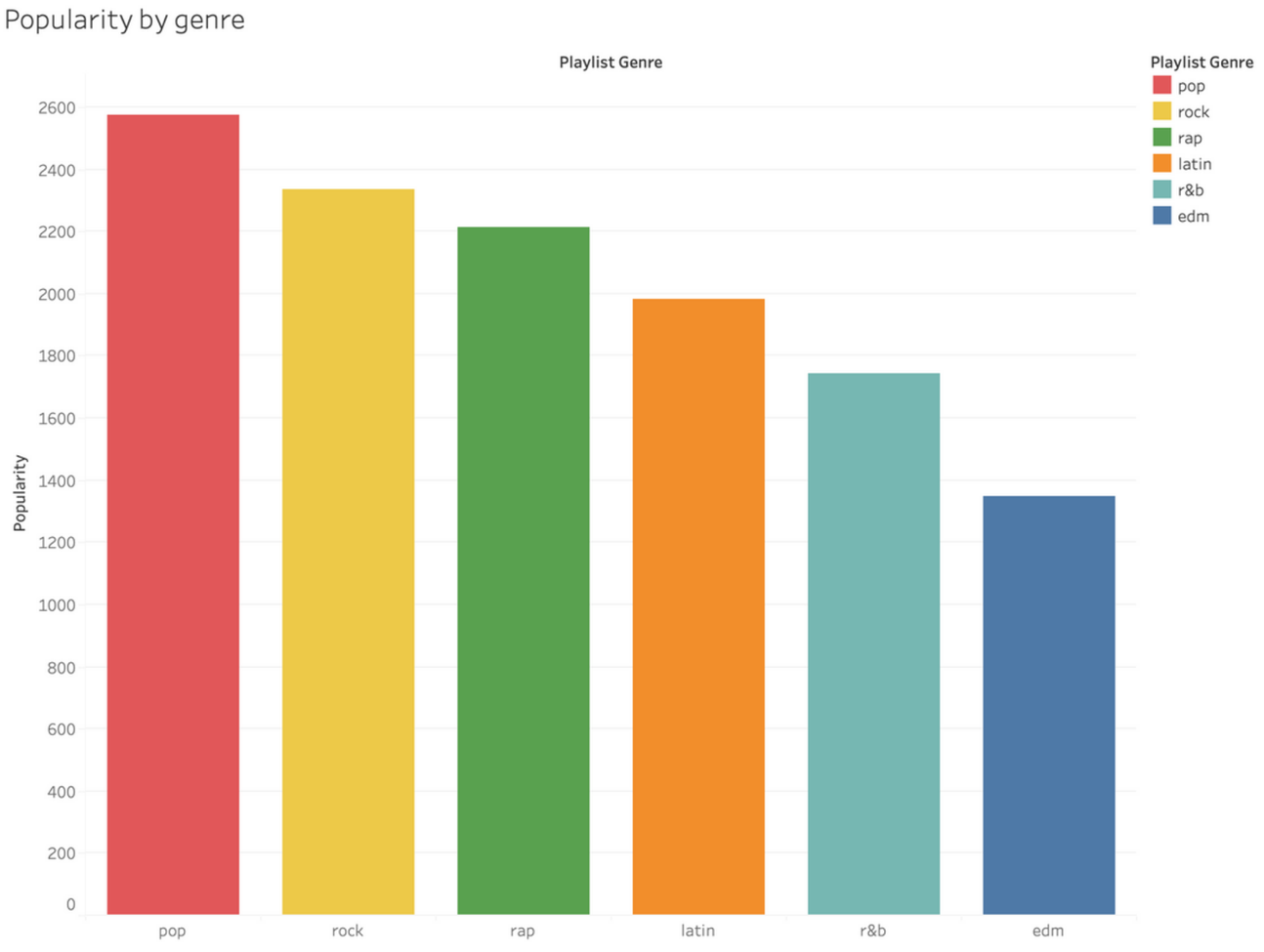
- As a music producer it is very crucial to keep a good track on the most popular artist for each genre and investing on them would be a wise choice to make money.
- David Guetta has most number of popular songs in edm category



Analyzing the most popular GENRE

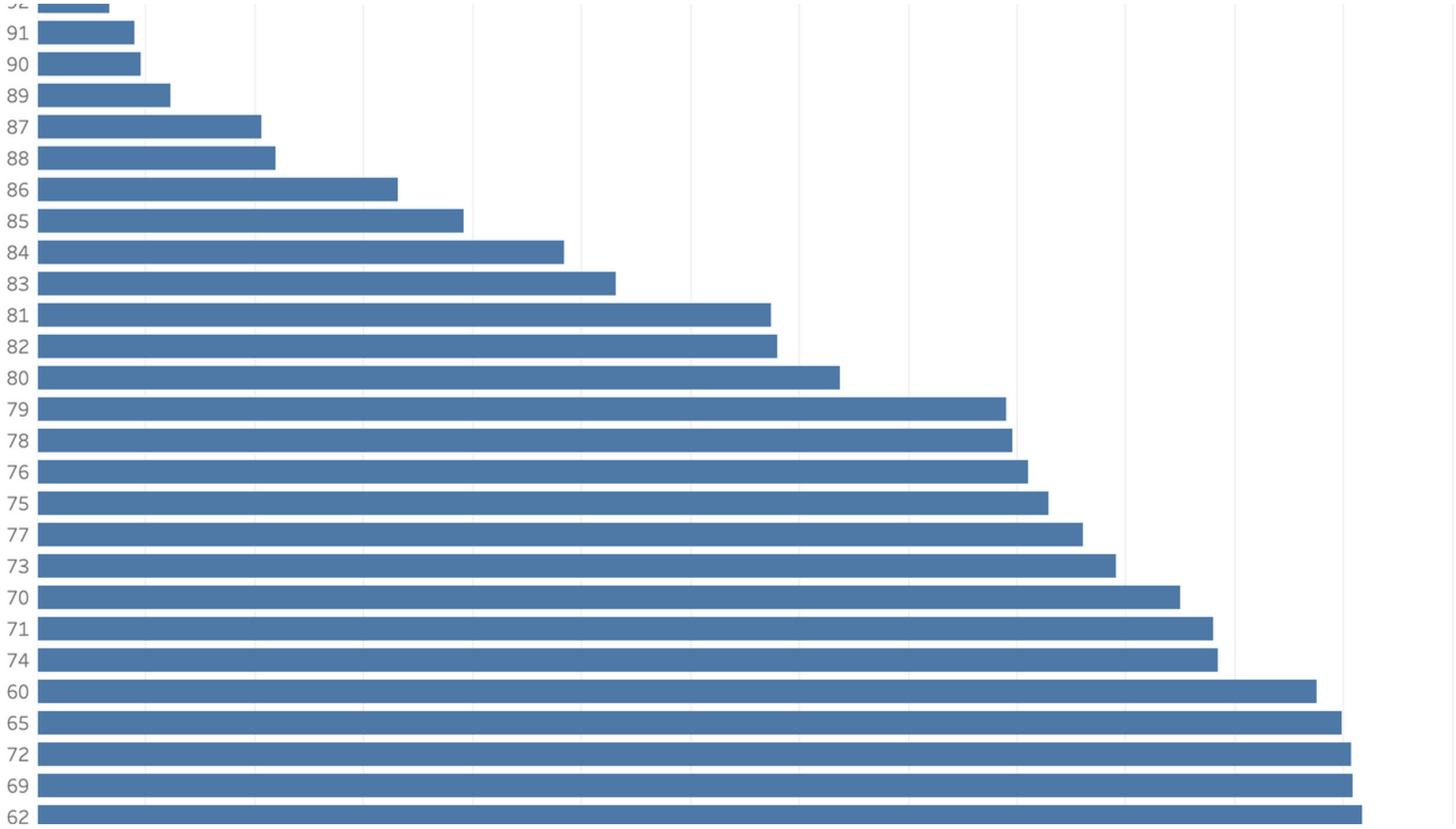
Pop is the most popular genre followed by rock

Investing on the most popular genre according to the trend could again be a wise choice.

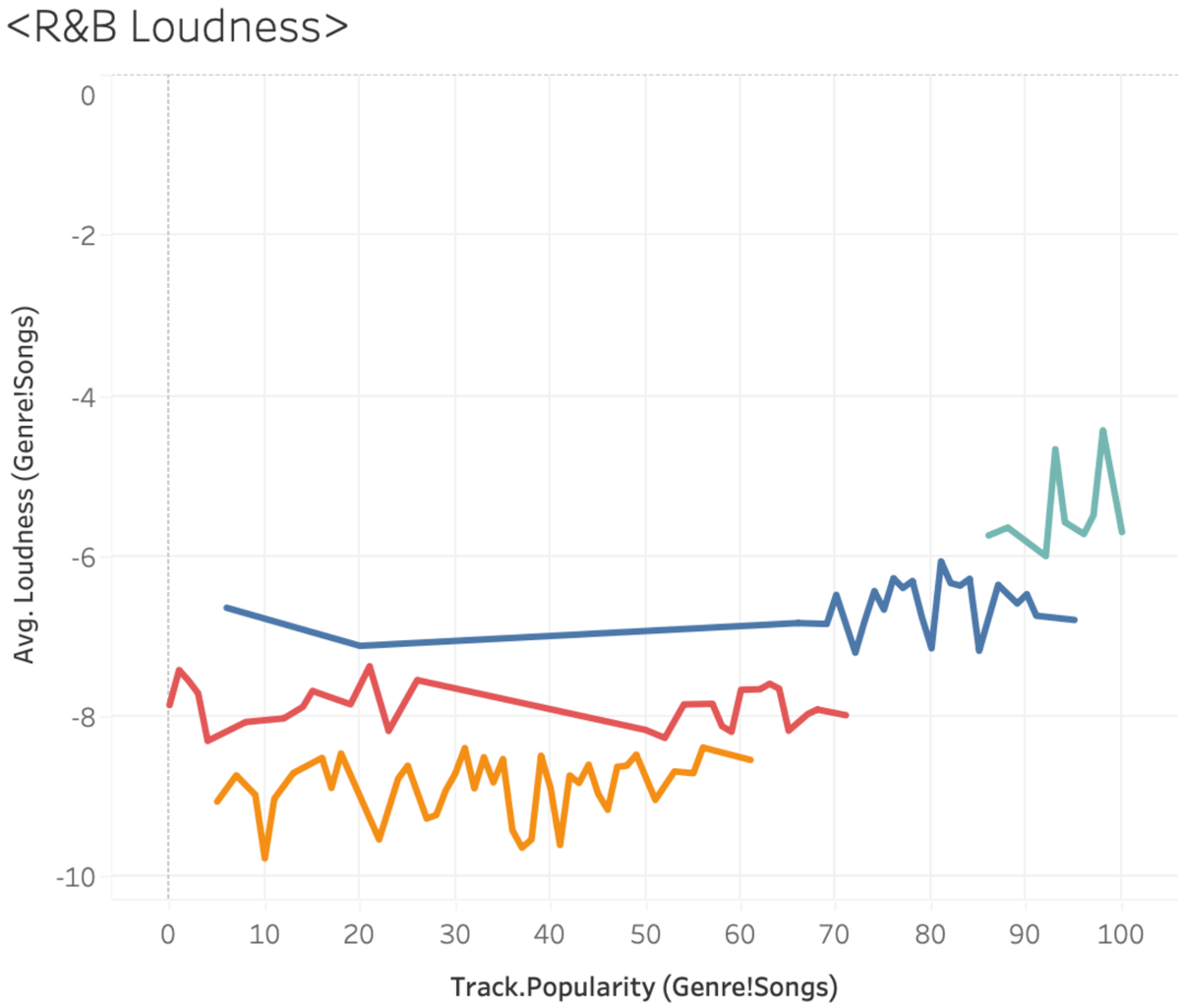


Duration vs popularity

- Most of the popular songs have good amount of duration.

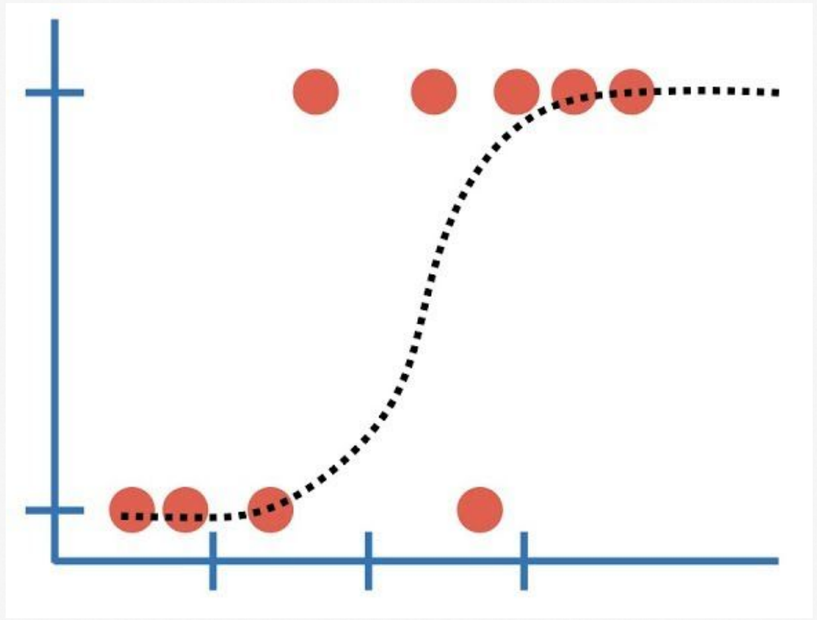


R&B Loudness



- THE TREND SHOWS THAT MORE POPULAR R&B SONGS TEND TO HAVE HIGHER LOUDNESS

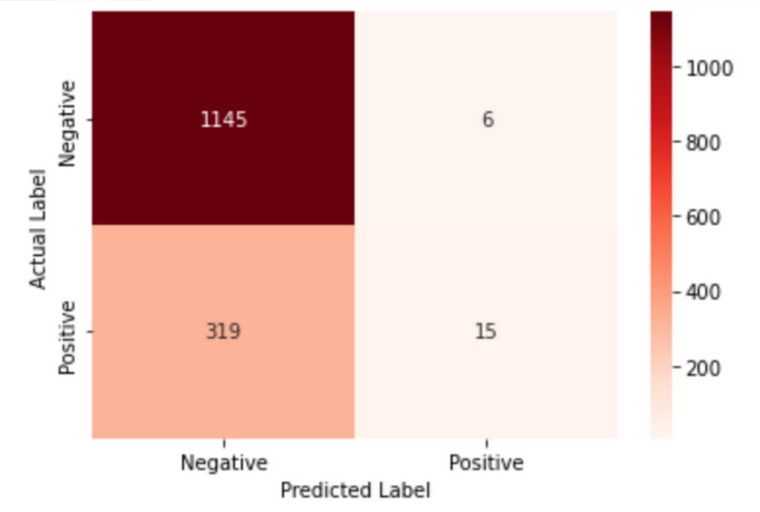
MODEL CHOICES AND REASONING



- Logistic regression is a good choice when dealing with binary classification problems and the relationship between input variables and the target variable is linear. It provides interpretable results and works well with small to medium-sized datasets.

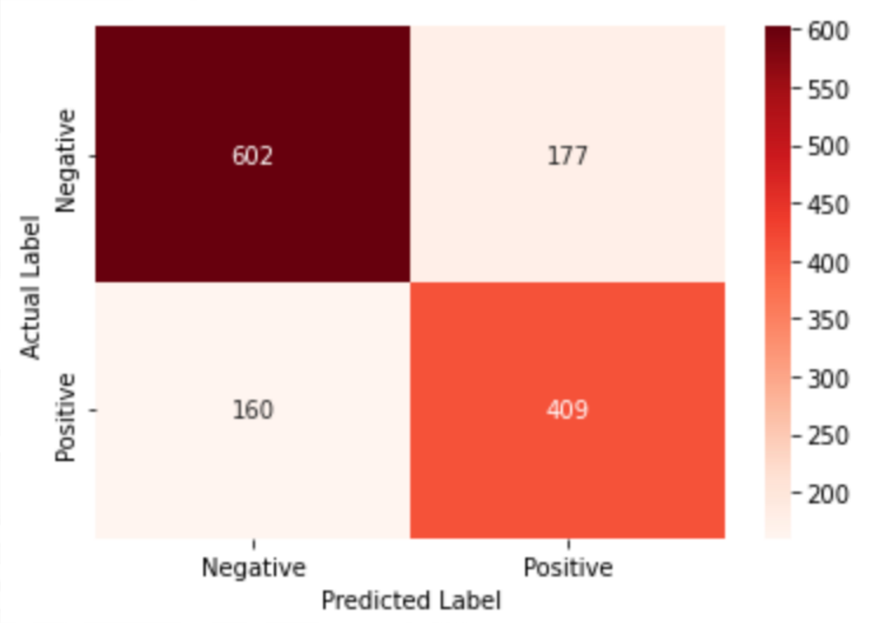
- XGBoost is a good choice when dealing with large and complex datasets, and when feature importance is a concern. It can handle both classification and regression problems, and is especially useful when dealing with imbalanced datasets.

Logistic Regression



- Logistic regression performs poorly in classifying minority classes in an imbalanced dataset
- It tends to optimize for overall accuracy and may not capture the underlying distribution of the data.
- It MISCLASSIFIED minority samples and have a higher false negative rate.
- Result is detrimental in applications where correctly identifying the minority class is critical.

XG Boost



- XGBoost is good for classifying minority class in an imbalanced dataset
- It can effectively learn from the minority class by giving them more weight during training
- It can handle non-linear relationships and complex interactions between features.
- XGBoost has a built-in regularization to prevent overfitting.
- which is especially important in imbalanced datasets with sparse samples.

Model Evaluation

Model	Genre					
	rap	edm	pop	rock	r&b	latin
Logistic	0.665	**0.784**	0.594	0.624	0.708	0.689
Xgboost	0.723	0.737	0.731	**0.75**	0.748	0.748

Recommendation

1. The model is built with hyperparameter tuning for getting very good accuracy. The music producers can take advantage of our model to make their decision about the release of the song.
2. The familiarity of the artist has a correlation to the popular songs. Having a popular artist increase the chance that the song could be a hit.
3. Feature analysis shows that ARTIST, SPEECHINESS, DURATION, DANCEABILITY, LOUDNESS and LIVENESS play a very important role in determining the success of a song. We need to consider different factors like loudness of the songs. Especially in R&B, the song maker would consider the songs with higher loudness.
4. We need to analyze more features to improve our model accuracy such as: region-based analysis, types of marketing campaign and so on.

THANKYOU